Gaussian Transforms Modeling and the Estimation of Distributional Regression Functions

Richard Spady and Sami Stouli

Nuffield College, Oxford and Johns Hopkins; University of Bristol

July 13, 2020

Introduction

- The modeling and estimation of conditional distribution functions are important for the analysis of various econometric and statistical problems.

- For instance, conditional distributions are core building blocks in
  - the identification and estimation of nonseparable models with endogeneity (e.g., Imbens and Newey, 2009; Chernozhukov, Fernandez-Val, Newey, Stouli and Vella, 2020, *Quantitative Economics*);
  - counterfactual distributional analysis (e.g., DiNardo, Fortin, and Lemieux, 1996; Chernozhukov, Fernandez-Val, and Melly, 2013).

- Conditional distributions are also a fruitful starting point for the formulation of general estimation methods (Spady and Stouli, 2018, *Biometrika*).

Introduction

- Consider a **continuous** outcome Y and a vector of covariates X.
- We observe that an objective function that characterizes $e = H(Y, X)$ such that

  (i) $e \sim N(0, 1)$,
  (ii) independent of X, and
  (iii) $y \mapsto H(y, X)$ is strictly increasing w.p.1,

  provides a valid characterization of the **'distributional regression functions'**

$$F_{Y|X}(Y \mid X) = \Phi\left(H(Y, X)\right)$$
$$Q_{Y|X}(u \mid X) = H^{-1}(\Phi^{-1}(u), X), \quad u \in (0, 1)$$
$$f_{Y|X}(Y \mid X) = \phi(H(Y, X))\frac{\partial H(Y, X)}{\partial Y}.$$

  where $\Phi(\cdot)$ is the Gaussian cumulative distribution function (CDF).

Introduction

- Consider a **continuous** outcome Y and a vector of covariates X.
- We observe that an objective function that characterizes $e = H(Y, X)$ such that

  (i) $e \sim N(0, 1)$,
  (ii) independent of X, and
  (iii) $y \mapsto H(y, X)$ is strictly increasing w.p.1,

  provides a valid characterization of the **'distributional regression functions'**

$$F_{Y|X}(Y \mid X) = \Phi\left(H(Y, X)\right)$$
$$Q_{Y|X}(u \mid X) = H^{-1}(\Phi^{-1}(u), X), \quad u \in (0, 1)$$
$$f_{Y|X}(Y \mid X) = \phi(H(Y, X))\frac{\partial H(Y, X)}{\partial Y}.$$

  where $\Phi(\cdot)$ is the Gaussian cumulative distribution function (CDF).

- **These distributional regression functions are known fnals. of $H(Y, X)$.**

Introduction

- From this observation we draw two themes:

1. **Modeling:**

   Working in terms of *Gaussian Transform Representations*, $e = H(Y, X)$, that satisfy properties **(i)-(iii)**.

2. **Objective function:**

   Formulate an objective function that characterizes the specified $H(Y, X)$ and preserves its properties, in particular monotonicity.

Contribution: Theory

- We formulate flexible models for **Gaussian Transform Representations**, $e = H(Y, X)$, as linear combinations of known functions of Y and X.
- We give an **ML characterization** of these representations, where the objective is concave and rules out nonmonotone solutions.
- We establish **existence and uniqueness** of the corresponding pseudo-true representations under misspecification.
- The resulting distributional models are then **KLIC optimal** approximations to the true data probability distribution (White, 1982).
- These approximations satisfy the **monotonicity** property of conditional CDFs by construction.

Contribution: Estimation

- We give **asymptotic properties** of the corresponding **MLE**.
- We extend the method to **adaptive Lasso** (Zou, 2006) to allow for model selection.
- We derive **asymptotic properties** of the corresponding estimators for **distributional regression functions**.
- For both MLE and adaptive Lasso we derive the corresponding **dual likelihood formulation** for implementation.

Agenda

1. **Gaussian Transforms Modeling**

2. **Maximum Likelihood Characterization**

3. **Estimation and Implementation**

4. **Empirical illustration**

Gaussian Transforms Modeling

- Throughout, we consider a **continuous** outcome random variable Y and a vector of explanatory variables X.

- The **Gaussian transform representation** for the CDF of $Y \mid X$,

$$H(Y, X) \equiv \Phi^{-1}\left(F_{Y|X}(Y \mid X)\right),$$

  is a zero mean and unit variance Gaussian random variable, and is independent from X (by construction).

- With $y \mapsto F_{Y|X}(y|X)$ strictly increasing, the corresponding map $y \mapsto H(y, X)$ is **strictly increasing** also.

Gaussian Transforms Modeling

- Let $W(X)$ and $S(Y)$ be vectors of **known** functions of X and Y, respectively. Denote the **derivative** of $S(Y)$ by $s(Y)$.
- We specify

$$
\begin{aligned}
H(Y,X) &= b_0' T(X,Y), \quad T(X,Y) = W(X) \otimes S(Y) \\
\frac{\partial H(Y,X)}{\partial Y} &= b_0' t(X,Y), \quad t(X,Y) = W(X) \otimes s(Y).
\end{aligned}
$$

- $H(Y,X)$ here is a **linear** combination of the dictionary elements, and the derivative is a **linear** combination of the derivative dictionary.

Gaussian Transforms Modeling

- Let $W(X)$ and $S(Y)$ be vectors of **known** functions of X and Y, respectively. Denote the **derivative** of $S(Y)$ by $s(Y)$.

- We specify

$$\begin{aligned} H(Y,X) &= b_0' T(X,Y), \quad T(X,Y) = W(X) \otimes S(Y) \\ \frac{\partial H(Y,X)}{\partial Y} &= b_0' t(X,Y), \quad t(X,Y) = W(X) \otimes s(Y). \end{aligned}$$

- $H(Y,X)$ here is a **linear** combination of the dictionary elements, and the derivative is a **linear** combination of the derivative dictionary.

- A **dictionary** $T(X,Y)$ always contains the elements $(1,Y)$, with corresponding elements $(0,1)$ for $t(X,Y)$.

- **Simplest specification** takes $S(Y) = (1,Y)'$ and $W(X) = (1,X)'$.

- '**Spline-Spline model**': an example of a **flexible specification** includes spline transformations both of X and of Y.

Gaussian Transforms Modeling

- The corresponding **density function** of $Y \mid X$ is

$$f_{Y|X}(Y \mid X) = \phi(H(Y,X))\frac{\partial H(Y,X)}{\partial Y} = \phi(b_0' T(X, Y))\{b_0' t(X, Y)\},$$

and the **log-density** is:

$$\begin{aligned}\log f_{Y|X}(Y \mid X) &= -\frac{1}{2}\left(\log(2\pi) + H(Y,X)^2\right) + \log\left(\frac{\partial H(Y,X)}{\partial Y}\right)\\ &= -\frac{1}{2}\left(\log(2\pi) + \{b_0' T(X, Y)\}^2\right) + \log(b_0' t(X, Y)).\end{aligned}$$

- This expression can then be used to fomulate an ML characterization of b_0, and hence of $H(Y,X)$ and the corresponding distributional regression functions.

Maximum Likelihood Characterisation (Population)

- Given our formulation, the **population ML objective** is:

$$Q(b) \equiv E\left[-\frac{1}{2}\left(\log(2\pi) + \{b'T(X,Y)\}^2\right) + \log\left(b't(X,Y)\right)\right]$$

- The corresponding **first- and second-derivative** functions are

$$\frac{\partial Q(b)}{\partial b} = E\left[-T(X,Y)\{b'T(X,Y)\} + \frac{t(X,Y)}{b't(X,Y)}\right]$$
$$\frac{\partial^2 Q(b)}{\partial b \partial b'} = -E\left[T(X,Y)T(X,Y)' + \frac{t(X,Y)t(X,Y)'}{\{b't(X,Y)\}^2}\right],$$

where $b't(X,Y) > 0$.

Notes/Interpretation for the ML problem

$$Q(b) = E\left[-\frac{1}{2}\left(\log(2\pi) + \{b'T(X,Y)\}^2\right) + \log\left(b't(X,Y)\right)\right].$$

- The true parameter vector b_0 maximizes $Q(b)$.
- The objective introduces a natural **logarithmic barrier function** in the form of the log of the Jacobian term.
- Thus the **monotonicity** requirement is imposed directly in the objective.
- The log Jacobian term is important also because it ensures **existence** of a maximiser under potential misspecification.
- When $E[T(X,Y)T(X,Y)']$ is nonsingular, the Hessian is negative definite so that $Q(b)$ is concave and has a **unique** maximizer.

Gaussian Transform Regression Theory: Summary

Model

A **Gaussian transform regression model** takes the form

$$H(Y,X) = b_0' T(X,Y) \mid X \sim N(0,1), \quad T(X,Y) \equiv W(X) \otimes S(Y), \qquad (1)$$

with derivative

$$\frac{\partial H(Y,X)}{\partial Y} = b_0' t(X,Y) > 0, \quad t(X,Y) \equiv W(X) \otimes s(Y). \qquad (2)$$

Regularity conditions

1. $E[||T(X,Y)||^2] < \infty$, $E[||t(X,Y)||^2] < \infty$, and the smallest eigenvalue of $E[T(X,Y)T(X,Y)']$ is bounded away from zero.

2. $f_{YX}(Y,X)$ is bounded away from zero with probability one.

Gaussian Transform Regression Theory: Summary

Theorem 1:

For model (1)-(2), $Q(b)$ has a unique maximum at b_0.

Theorem 2:

There exists a unique maximum b^* to $Q(b)$.

Theorem 3:

The pseudo-true density $f^*_{Y|X}(Y \mid X) \equiv \phi(T(X,Y)'b^*)\{t(X,Y)'b^*\}$ is the KLIC-closest approximation to $f_{Y|X}(Y \mid X)$ in the specified class of cond. density functions.

Connection with Distribution Regression Models

- Model (1)-(2) also arises from specifying $H(Y, X)$ as a linear combination of the known functions $W(X)$

$$H(Y, X) = W(X)'\beta(Y) \tag{3}$$

with $\beta(Y) = (\beta_1(Y), \ldots, \beta_K(Y))'$ a vector of random coefficients specified as

$$\beta_k(Y) = b_k' S(Y), \quad k \in \{1, \ldots, K\}, \quad K \equiv \dim(W(X)). \tag{4}$$

Connection with Distribution Regression Models

- Model (1)-(2) also arises from specifying $H(Y,X)$ as a linear combination of the known functions $W(X)$

$$H(Y,X) = W(X)'\beta(Y) \tag{3}$$

with $\beta(Y) = (\beta_1(Y), \ldots, \beta_K(Y))'$ a vector of random coefficients specified as

$$\beta_k(Y) = b_k' S(Y), \quad k \in \{1, \ldots, K\}, \quad K \equiv \dim(W(X)). \tag{4}$$

- Together (3)-(4) give the linear form

$$H(Y,X) = \sum_{k=1}^{K} W_k(X)\beta_k(Y) = \sum_{k=1}^{K} W_k(X)\{b_k' S(Y)\} = b_0'[W(X) \otimes S(Y)].$$

Connection with Distribution Regression Models

- Model (1)-(2) also arises from specifying $H(Y, X)$ as a linear combination of the known functions $W(X)$

$$H(Y, X) = W(X)'\beta(Y) \tag{3}$$

with $\beta(Y) = (\beta_1(Y), \dots, \beta_K(Y))'$ a vector of random coefficients specified as

$$\beta_k(Y) = b_k' S(Y), \quad k \in \{1, \dots, K\}, \quad K \equiv \dim(W(X)). \tag{4}$$

- Together (3)-(4) give the linear form

$$H(Y, X) = \sum_{k=1}^{K} W_k(X)\beta_k(Y) = \sum_{k=1}^{K} W_k(X)\{b_k' S(Y)\} = b_0'[W(X) \otimes S(Y)].$$

- For $H(Y, X) \mid X \sim N(0, 1)$, then

$$F(y \mid X) = \Phi(W(X)'\beta(y)), \quad y \in \mathcal{Y},$$

a (Gaussian) distribution regression model.

Maximum Likelihood Estimation

- Given a sample $\{(y_i, x_i)\}_{i=1}^n$, the ML objective function is

$$Q_n(b) = \sum_{i=1}^{n} \left\{ -\frac{1}{2}[\log(2\pi) + (b'T(x_i, y_i))^2] + \log(b't(x_i, y_i)) \right\}.$$

- The MLE is

$$\hat{b} \equiv \arg\max Q_n(b).$$

- Consistency and asymptotic normality of $\hat{b}$ follow from ML theory for concave objective function.

- Asymptotic distribution of distributional regression functions follows by the Delta method.

- This is a **convex programming problem**.

Maximum Likelihood Estimation: Adaptive Lasso

- For **model selection** and in order to allow for the dimension of $T(x_i, y_i)$ to be large (i.e., singularity and "$p < n$") the objective can be augmented with an adaptive Lasso penalty:

$$\hat{b}_{\mathrm{AL}} \equiv \arg\max Q_n(b) - \lambda_n \sum_{l=1}^{\dim(T(x_i,y_i))} \widehat{w}_l |b_l|,$$

- $\lambda_n > 0$ is a **penalization parameter** and the **weights** $\widehat{w}_l$ are defined as

$$\widehat{w}_l \equiv \begin{cases} \frac{1}{|\hat{b}_l|} & \text{if } \hat{b}_l \neq 0 \\ 0 & \text{if } \hat{b}_l = 0 \end{cases}, \quad l = 1, \ldots, \dim(T(x_i, y_i)).$$

- Asymptotic properties of $\hat{b}_{\mathrm{AL}}$ follow from adapt. Lasso theory under misspecification (e.g., Lu, Goldberg, and Fine, 2012)

- This is also a **convex programming problem**.

Implementation: Dual Likelihood Formulation

(i) The **dual likelihood problem** is

$$\min -n\left(\frac{1}{2}\log(2\pi)+1\right)+\sum_{i=1}^{n}\left\{\frac{u_i^2}{2}-\log\left(-v_i\right)\right\}$$

$$\text{subject to} \quad -\sum_{i=1}^{n}\{T(x_i,y_i)u_i+t(x_i,y_i)v_i\}=0 \qquad (5)$$

the dual Gaussian transform regression problem, with solution $\widehat{\alpha} \equiv (\widehat{u}', \widehat{v}')'$.

(ii) The program (5) admits the **method-of-moments representation**

$$\sum_{i=1}^{n}\left\{-T(x_i,y_i)\{b'T(x_i,y_i)\}+\frac{t(x_i,y_i)}{b'\,t(x_i,y_i)}\right\}=0,$$

the first-order conditions of the primal ML problem.

(iii) The **solutions** of the two problems are related by

$$\widehat{u}_i=\widehat{b}'\,T(x_i,y_i), \quad \hat{v}_i=-\frac{1}{\widehat{b}'\,t(x_i,y_i)}, \quad i=1,\ldots,n.$$

(iv) Strong duality, i.e., the value of the primal ML problem equals the value of (5).

Discussion

- Difficulties arise in the formulation of flexible models and in the choice of an objective function for the characterization of $F_{Y|X}(Y \mid X)$.
- Various formulations exist that feature advantages and drawbacks. E.g.,
  - **Quantile regression models** (Koenker and Bassett, 1978) specify the cdnal. quantile function as a linear combn. of known functions of X.
  - **Distribution regression models** (Foresi and Peracchi, 1995; Chern., Fernandez-Val, and Melly, 2013) specify the cond. CDF as a probability transform of a linear combn. of known functions of X.

Discussion

- Difficulties arise in the formulation of flexible models and in the choice of an objective function for the characterization of $F_{Y|X}(Y \mid X)$.

- Various formulations exist that feature advantages and drawbacks. E.g.,
  - **Quantile regression models** (Koenker and Bassett, 1978) specify the cdnal. quantile function as a linear combn. of known functions of X.
  - **Distribution regression models** (Foresi and Peracchi, 1995; Chern., Fernandez-Val, and Melly, 2013) specify the cond. CDF as a probability transform of a linear combn. of known functions of X.

- For both approaches, the corresponding objective function characterizes the object of interest pointwise.
  - As a result, the defining feature of monotonicity may not be preserved in finite samples and under mispecification (Chernozhukov, Fernandez-Val, and Galichon, 2010).

Discussion

- Another approach is to specify flexible models and an objective function that characterizes these models globally.

  - **Dual regression models** (Spady and Stouli, 2018) specify the quantile function as a linear combn. of known functions of both X and a stochastic element that satisfies the properties of a cond. CDF.

  - Dual regression solutions preserve **monotonicity**.

Discussion

- Another approach is to specify flexible models and an objective function that characterizes these models globally.

  - **Dual regression models** (Spady and Stouli, 2018) specify the quantile function as a linear combn. of known functions of both X and a stochastic element that satisfies the properties of a cond. CDF.

  - Dual regression solutions preserve **monotonicity**.

- The implied modeling of the cond. CDF is indirect. This is not innocuous.

- The method is not endowed with an ML interpretation.

Empirical Illustration

- We use a dataset gathering 3,650 consecutive **daily maximum temperatures** in Melbourne, y_t.
- We estimate conditional quantile functions (CQF) of y_t given y_{t-1} and the corresponding densities.
- This dataset was used by Koenker (2005) to illustrate nonlinear quantile regression, and originally analyzed by Hyndman, R.J., Bashtannyk, D.M. and Grunwald, G.K. (1996).
- This dataset is challenging because **the distribution of today's temp. varies across yesterday's temp. values**:
  - temperatures following very hot days are **bimodal**, with the lower mode corresponding to a 'break' in the temperature (i.e., a much cooler temperature).
  - The temperatures of days following 'normal' days are **unimodal**.
- This dataset allows for the illustration of the main features of each class of Gaussian transform representations.

Empirical Illustration

- We illustrate the main features of the following three specifications

  1. **Linear-X and Spline-Y specification**:
  $W(X) = (1, X)'$ and $S(Y)$ includes a vector of cubic spline functions.

  2. **Spline-X and Linear-Y specification**:
  $W(X)$ includes a vector of cubic spline functions and $S(Y) = (1, Y)'$.

  3. **Spline-Spline specification**:
  both $W(X)$ and $S(Y)$ include a vector of cubic spline functions.

- **Specification 1 & 2:** we estimate 9 models with 4 to 12 degrees of freedom (increasing sequence of equispaced knots).

- **Spline-Spline:** we estimate 18 models, with 4 and 5 degrees of freedom for splines in $S(Y)$ and 4 to 12 degrees of freedom for splines in $W(X)$.

- For each specification, select the model with **smallest BIC**.

CQFs: Linear-X, Spline-Y (BIC=19877)

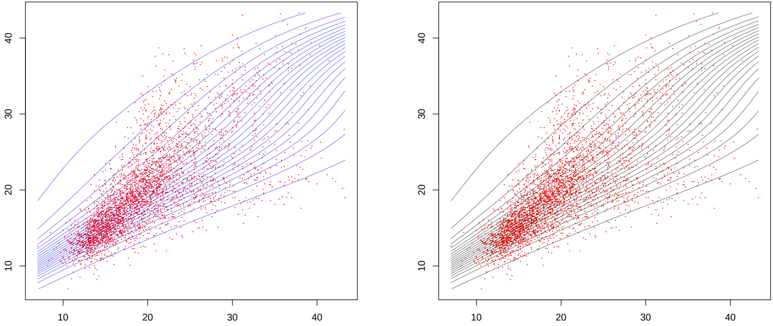


Figure: No penalization (left) and adaptive Lasso (right). Quantile grid: $(0.01, 0.05, 0.1, \ldots, 0.95, 0.99)$.

CQFs: Spline-X, Linear-Y (BIC=19425)

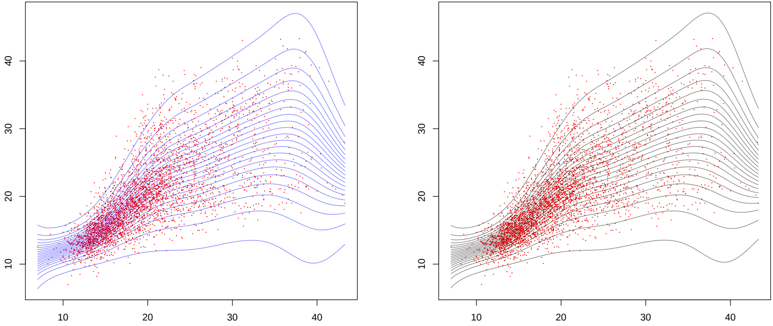


Figure: No penalization (left) and adaptive Lasso (right). Quantile grid: $(0.01, 0.05, 0.1, \ldots, 0.95, 0.99)$.

CQFs: Spline-Spline (BIC=19350)

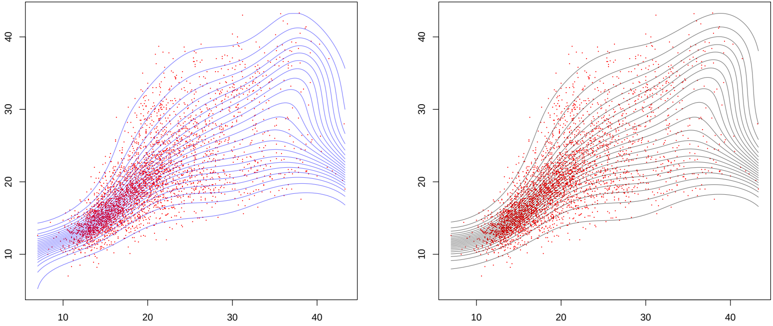
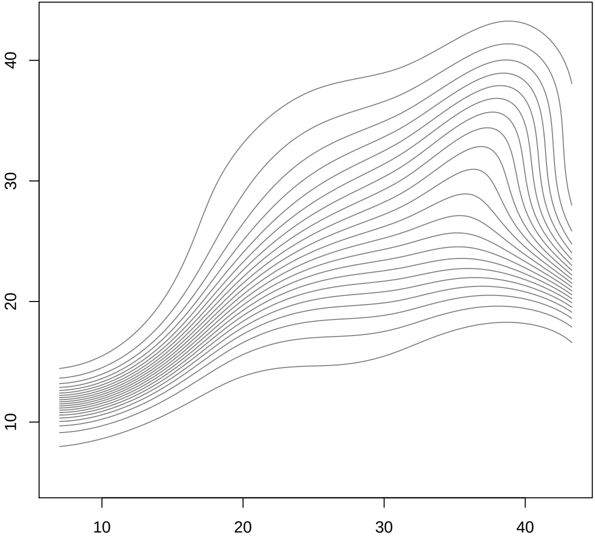
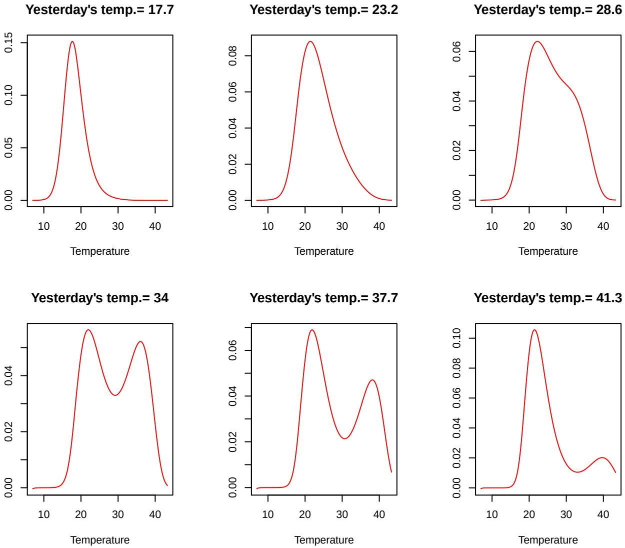


Figure: No penalization (left) and adaptive Lasso (right). Quantile grid: $(0.01, 0.05, 0.1, \ldots, 0.95, 0.99)$.

Melbourne via Spline-Spline - Adaptive Lasso



Conditional density functions



Conclusion

- Writing the estimation problem in the $e = H(Y, X)$ form is convenient.

- Allows for the joint formulation of representations and an objective function that preserve:
  1. nonseparability,
  2. monotonicity (both in finite-samples and under general mispecification),
  3. KLIC optimality, and
  4. closed-form modeling of the Gaussian transform ($\Rightarrow$ considerable computational simplification).

- Wide range of natural applications and extensions.